Title

Integrating the environmental and genetic architectures of mortality and aging

Authors

M. Austin Argentieri[1,2]*, Najaf Amin[1], Alejo J. Nevado-Holgado[3], William Sproviero[3], Jennifer A. Collister[1], Sarai M. Keestra[4,5], Aiden Doherty[1], David J. Hunter[1], Alexandra Alvergne[2,6], Cornelia M. van Duijn[1]*

Affiliations

1. Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
2. School of Anthropology and Museum Ethnography, University of Oxford, Oxford, United Kingdom
3. Department of Psychiatry, University of Oxford, Oxford, United Kingdom
4. Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, The Netherlands
5. Amsterdam Reproduction & Development, Amsterdam UMC, University of Amsterdam, The Netherlands
6. ISEM, University of Montpellier, CNRS, IRD, Montpellier, France

* Correspondence to M. Austin Argentieri, austin.argentieri@anthro.ox.ac.uk, and Cornelia M. van Duijn, cornelia.vanduijn@ndph.ox.ac.uk

One Sentence Summary

The exposome is a major mortality determinant irrespective of genetic disease risk via shaping distinct biological and multimorbidity patterns.

Abstract

In this study, we aimed to quantify the relative roles of the genome and exposome in aging and mortality. We conducted an exposome-wide analysis in the UK Biobank (n=492,567) using independent discovery and replication sets to systematically identify exposures that are associated with mortality (median 12.5 follow-up years), 25 age-related diseases, 25 aging biomarkers, and 3 cardiometabolic risk factors. We identified 41 independent exposures that were associated with mortality, most of which were associated with consistent biological and multimorbidity signatures. Compared with a simple model composed of age and sex, polygenic risk scores for 22 major causes of death and aging phenotypes explained an additional 2% of mortality variation, whereas the exposome explained an additional 19% of variation. While genetics explained a high proportion of variation in dementias, breast, ovarian, and colorectal cancers, the exposome explained a high proportion of variation for diseases of the lung, heart, and liver. Our findings provide a comprehensive map of the relative contributions of environment and genetics to mortality and common age-related diseases, and suggest that environment-focused interventions are likely to have the highest impact on ameliorating premature mortality.

Main Text

Human aging is a complex process that initially manifests as sub-clinical and biological changes that begin to accumulate from mid-life onwards (*1-3*). These systemic biological changes are major drivers of common age-related diseases (*4-6*) and disease multimorbidity (*7, 8*), which in turn are the major causes of premature mortality worldwide (*9*). While there have been major advancements in understanding the complex genetic etiology of age-related diseases, genetic studies show only a modest effect of the genome on lifespan (*10, 11*). Instead, the nearly twofold increase in global human lifespan during the past 200 years (*12*) has been largely attributed to changes in human environments (*13*). Epidemiological research has made progress in relating individual environmental and behavioral exposures to age-related diseases and mortality, however few studies have comprehensively examined the exposome (i.e., the total set of interrelated environmental exposures throughout the life course) in relation to these outcomes (*14, 15*). In the field of genetic epidemiology, the use of genome-wide approaches have greatly increased the positive predictive value (*16*) and reproducibility (*17*) of findings. Transitioning from studies of single environmental exposures to exposome-wide studies will provide similar advancements.

Building upon the framework of genomic research on complex diseases, we conducted an exposome-wide analysis using data from the UK Biobank (n=492,567) to systematically identify exposures associated with mortality and multiple stages of the aging process, as captured by multimorbidity of common age-related disorders and their risk factors. We first systematically identified exposures that reliably associate with the most critical outcome of all age-related diseases – mortality. We then tested exposures associated with mortality in relation to: (i) incidence of age-related diseases that are either major causes of death or highly prevalent in aging populations (25 total); (ii) cross-sectional patterns of all age-related blood biomarkers available in the UK Biobank (25 total); and (iii) prevalence of three major cardiometabolic risk

factors (obesity, hypertension, dyslipidemia). We then examined the genetic and environmental architecture of mortality and age-related diseases to quantity the relative contribution of the exposome vs. polygenic risk to explaining variation in each outcome.

Mortality and age-related disease rates

The final study sample included 492,567 UK Biobank participants (Fig. 1). Exposome-wide association study (XWAS), aging biomarker, and chronic disease analyses were carried out using UK Biobank participants recruited in England (n=436,891). Participants recruited in Scotland/Wales (n=55,676) were held out as a validation set used only to validate final multivariable disease models. There were 31,716 deaths from all causes among English participants after a median 12.5 years of follow up (Table S1). The majority (74.5%) of deaths were premature deaths (i.e., occurring before 75 years of age; Fig. 2a) and 75% of deaths occurred in those who were overweight or obese with a body mass index (BMI) ≥ 25 kg/m^2 (Fig. 2b). Women had a lower rate of all-cause mortality compared with men (5.4% in women vs 9.4% in men; Table S1). Among Scottish/Welsh participants, there were 5,267 deaths from all causes after a median 13.7 years of follow up (Table S3). Mortality by cause of death for participants recruited both in England and Scotland/Wales is given in Tables S4-S5.

The number incident cases for all age-related diseases studied in English participants ranged from 856 (brain cancer) to 45,879 (osteoarthritis), as shown in Fig. 2c and Table S6; summary statistics for all cross-sectional outcomes (3 cardiometabolic risk factors, 25 baseline aging biomarkers) are given in Tables S6-S7. Key demographic prevalence rates for English participants are shown in Fig. 2d.

Exposome-wide analysis of mortality

XWAS of all-cause mortality were conducted by serially testing each exposure in relation to mortality via Cox proportional hazards models using independent discovery and replication subsets of the UK Biobank study population (Fig. 1). No significant differences were observed in XWAS regression betas when calculated separately in women and men (Fig. 3a). In a final mortality XWAS using pooled data from both women and men, 110/164 exposures (67.1%) were significantly replicated (Fig. 3b). Smoking, renting public housing (compared with home ownership), Townsend deprivation index, and living with a partner were the exposures most strongly associated with mortality. Sensitivity analyses excluding participants who died within the first 4 years of follow up and testing interactions between each exposure and a baseline poor health indicator suggested that largely there is no strong statistical evidence for reverse causation bias in our XWAS results (Fig. S1-S2), with only 15 exposures identified whose associations with mortality were likely completely explained by prevalent disease status (Fig. S1). These exposures were discarded, leaving 95/110 (86.4%) exposures whose associations with mortality were not likely explained by pre-existing morbidity. Summary statistics from all mortality XWAS analyses are given in Supplementary Files SF3-SF7.

PheWAS of replicated exposures

Each exposure replicated in the mortality XWAS and passing the above sensitivity analyses was checked for possible collinearity with other exposures and mismeasurement by conducting a phenome-wide association study (PheWAS) where the replicated exposure was treated as the outcome variable and regressed against all baseline phenotypes present in the UK Biobank using either logistic or linear regression (Supplementary Information). Using this method, we detected a further 10 exposures that associated extremely strongly with either: (i)

disease, frailty, or disability phenotypes; or (ii) another exposure such that it likely does not represent new information. For example, we found that never eating sugar was very strongly associated with having a diabetes diagnosis (OR: 4.2, $p < 3.5 \times 10^{-154}$) and insulin medication intake (OR: 5.43, $p < 6.3 \times 10^{-127}$), indicating that this exposure is mostly capturing diabetes and disability-related dietary restriction (Fig. S4). We also found that the number of vehicles in a participant's household was very strongly associated with household income (Beta: 1.05, $p < 8.1 \times 10^{-12}$), while inversely associated with living in council housing vs. home ownership (Beta: -0.98, $p < 5.01 \times 10^{-56}$) and being unemployed due to a disability (Beta: -0.62, $p < 1.4 \times 10^{-245}$). These findings indicate that this exposure is mostly capturing socioeconomic and disability status (Fig. S5). Exposures showing mismeasurement from PheWAS were discarded and not carried forward to further analyses, leaving 86 remaining exposures. Summary statistics from all PheWAS are given in Supplementary Files SF62-SF108 and interactive volcano plots for each are shown in the Online Materials.

Cluster architecture of the exposome

We observed high degrees of correlation between exposures replicated in the XWAS (90% of variable pairs had a significant Bonferroni-corrected correlation p-value below 0.001), indicating that some mortality associations observed in the XWAS may be confounded due to this correlation structure. To address this, we used hierarchical clustering to organize replicated exposures that passed all sensitivity analyses into 9 unique clusters (Fig. 3c). We first conducted multivariable mortality Cox models within each cluster by adding all exposures from the cluster into a single Cox model (Fig. S11a). We discarded exposures that did not pass multicollinearity tests or were not significant in this within-cluster model. We then grouped the remaining significant cluster exposures into two large superclusters (clusters 1-5 and 6-9) and conducted a Cox mortality model within each of these superclusters to also account for long-

range correlation confounding. Using this method, we identified 41 exposures that remained significant in these cluster multivariable models (Fig. S11b).

We found that multiple dimensions of physical activity (general physical activity [IPAQ]; leisure time physical activity [LTPA]; total sedentary time) all exhibit significant, independent effects on mortality, even when included in the same model and adjusted for a wide number of other correlated variables. To further interrogate the robustness of this result, we also tested the association between total physical activity and all-cause mortality using accelerometer data available for a subset of participants (n=85,520; Supplementary Information). Self-reported physical activity variables underestimated the effect of accelerometer average on mortality by only 3% as measured by R^2 (Table S13).

Patterns of age-related multimorbidity and biological mechanisms

To test whether the 41 identified exposures were associated with not just mortality but multiple stages of the aging process, we tested each exposure individually in relation to incidence of 25 age-related diseases via Cox proportional hazards models (8-15 years of follow-up), as well as to cross-sectional patterns of 25 age-related biomarkers and 3 cardiometabolic risk factors (hypertension, obesity, dyslipidemia) via linear and logistic regression. Each of the 41 exposures independently associated with mortality was associated with a wide range of aging biomarkers that span diverse organ systems and mechanisms (Fig. 4a). On average, each exposure was associated with 21.9 biomarkers (out of 25). Four exposures were associated with all 25 biomarkers (Townsend deprivation index, smoking status, hours of sleep, ethnicity) and nine with 24/25 biomarkers in total (unenthusiasm frequency, total sedentary time, home ownership, taking multivitamin supplements, LTPA, household income, years of education, relative body size at 10 years old, alcohol intake frequency). Summary statistics from all biomarker analyses are given in Supplementary Files SF8-SF32. Cardiometabolic risk factors

studied were cross-sectionally associated with nearly every exposure studied (Fig. 4b). Notably, hypertension was cross-sectionally associated with all exposures tested.

Each of the 41 exposures independently associated with mortality was also prospectively associated with concurrent incidence of multiple chronic diseases (Fig. 4c), indicating that each exposure is a likely catalyst of disease multimorbidity. On average, each exposure was associated with 11.7 chronic diseases (out of 25). Smoking was associated with 21 (current smoking status) and 20 diseases (pack years). Household income and home ownership were associated with 19/25 diseases, followed by tiredness frequency and Townsend deprivation index (18/25), and IPAQ physical activity group (17/25). Of note, we found no associations between any exposure and incidence of brain cancer. We generally observed prostate cancer associations in the opposite direction as other diseases for the same exposure, however we performed further sensitivity analyses (Supplementary Information) to rule out any detection bias underlying these results for key exposures such as smoking (Table S14). Summary statistics from all chronic disease and cardiometabolic risk factor analyses are given in Supplementary Files SF33-SF60.

For approximately 75% of exposures, we found remarkable consistency in the direction of effect when comparing associations across mortality, aging biomarker, and disease analyses. For example, exposures associated with decreased mortality risk (employment, household income, education, living with a partner, IPAQ, LTPA, gym use, glucosamine and fish oil supplements, cereal fiber intake) were associated with decreased risk of nearly two thirds of all chronic diseases studied; decreased risk of obesity, dyslipidemia, and hypertension; increased levels of biomarkers indicating better health (vitamin D, HDL cholesterol, phosphate, albumin) or slower aging (LTL, IGF-1); and decreased levels of detrimental biomarkers (C-reactive protein, blood glucose, hbA1c, cystatin C, triglycerides). Exposures associated with increased mortality risk (smoking, lack of home ownership, being unemployed, deprivation, air pollution, maternal smoking around birth, sedentary lifestyle, poor mental health, poor sleep, processed and red

meat intake) were generally associated with opposite patterns of these same diseases, risk factors, and biological mechanisms.

Environmental and genetic architectures of mortality and age-related diseases

To determine the contribution of age and sex, exposome, and genetic risk in describing variation in mortality and each of the studied age-related diseases, we calculated stepwise multivariable Cox models beginning with just age and sex (model 1), then adding either polygenic risk information (model 2) or exposome (i.e., all exposures associated with the outcome; model 3), and finally adding both the exposome and PRS together (model 4). Models were first calculated among participants recruited in England (n=436,891) and then validated in the independent set of participants recruited in Scotland/Wales (n=55,676).

Compared with a model containing age and sex, we found that adding polygenic risk scores (PRS) for 22 diseases that are either major causes of death or aging phenotypes only increased the total mortality model R^2 by 2% (Fig. 5a; Tables S15-S16). By contrast, we found that adding all 41 exposures associated with mortality (i.e., exposome) to age and sex increased the total mortality model R^2 by 17-19% (model 3 vs 1), and that adding the 41 exposures to the model with age, sex, and all PRS increased the total mortality model R^2 by 16-17% (model 4 vs 2). While the combined effect of the exposome explained a large proportion of mortality variation, we found that individually most exposures only explained a small proportion of total mortality variation (Fig. 3d). Effect estimates for all 41 exposures in model 3 are shown in Fig. S11c. Variable importance plots for other diseases studied are shown in Fig. S11-S19. Effect estimates for all exposures in model 3 for all age-related diseases studied are shown in Fig. S20-44.

Models including age and sex, exposome, and PRS captured >50% of variation in most outcomes studied, with the exception of colorectal cancer, pancreatic cancer, leukemia, breast

and ovarian cancers, and osteoarthritis (Fig. 5a). For all-cause mortality and all age-related diseases studied, the relative importance of age, sex, exposome, and PRS in the training multivariable models (English participants) are shown in Fig. 5b according to the proportions of the total model chi-squared (X^2) that each variable category explained in model 4. The exposome explained the most amount of model variation for lung cancer, emphysema/COPD, chronic liver diseases, and rheumatoid arthritis. Certain outcomes seem to be more influenced by polygenic risk than the exposome, such as breast and prostate cancers, Alzheimer's disease (AD), all-cause dementia, macular degeneration, and colorectal cancer. Of note, ovarian cancer and type 2 diabetes showed a smaller contribution of age and sex, with the exposome and PRS explaining the majority of variation in equal parts. Lastly, a number of outcomes showed age and sex as the most influential determinants, but also showed the exposome explaining the majority of the residual variation not explained by age and sex. These include all-cause mortality, esophageal cancer, ischemic heart disease, and cerebrovascular diseases.

Discussion

This study provides the first comprehensive assessment of the relative contributions of genetic and environmental influences to aging. We show that the exposome explains a large percentage of mortality variation beyond the contribution of age, sex, and polygenic disease risk. We further demonstrate that the exposome shapes mortality risk through influencing a wide range of aging biological mechanisms and multimorbidity. For most age-related disorders studied, we find that the effect of the exposome exceeds that of the genome. We also find that exposome risk is composed of many interrelated factors that individually may have small effects, but when combined additively explain a substantial amount of variation for mortality and certain age-related diseases such as emphysema/COPD, lung cancer, chronic liver diseases, and rheumatoid arthritis.

Our results provide critical insights to address gaps in knowledge about the environmental influences of biological aging mechanisms (*2*), and further demonstrate that many age-related diseases share a common environmental etiology that drives disease multimorbidity and ultimately premature mortality. We found that generally diseases fell into two large blocks. The first included cardiometabolic diseases, lung and kidney disease, dementia, vascular dementia, osteoporosis, osteoarthritis, and rheumatoid arthritis, all of which showed a shared environmental etiology involving the majority of exposures associated with mortality. The second block of disease included most common cancers and the common neurodegenerative disorders (AD, Parkinson's disease, macular degeneration), which were associated with fewer environmental determinants but still showed similar direction of effects as seen in the first block.

We found that the impact of the genome is relatively high for colorectal, breast, ovarian, and prostate cancers, as well as all-cause dementia and AD. Of note, although we saw associations between known determinants of colorectal cancer such as physical activity, smoking, and intakes of cereal fiber, processed meat, and red meats, we found that the contribution of these exposures to explaining colorectal cancer variation was limited compared to genetic risk. Conversely, we show the impact of the genome to be relatively low and the impact of the exposome to be high for lung cancer and emphysema/COPD, as well as chronic liver and kidney diseases, ischemic heart disease, cerebrovascular diseases, rheumatoid arthritis, osteoarthritis.

Our study design improves substantially upon XWAS and exposome analyses published to date (see our systematic review in the Supplementary Information) by conducting cluster analyses to detect correlation confounding and PheWAS to detect mismeasurement in exposures. When compared with the only previously published "environment-wide" analysis of mortality (*18*) that focused on a narrower range of chemical and lifestyle exposures in a small sample (n=6,008), our study identified approximately 17x more factors associated with all-cause mortality and improved the final variance explained (mortality R^2) by 31x from 2.1% in this

previous study to 66%. This demonstrates the importance of using large discovery datasets and testing as broad a range of exposures as possible in exposome research.

Throughout our exposome analyses, measures of physical activity, smoking, and individual socioeconomic status (household income and home ownership) had the strongest effects on mortality and were associated with the greatest number of age-related diseases and aging biomarkers. While numerous previous studies have already documented the significant roles of smoking, physical activity, and socioeconomic inequality (e.g., income, employment, education) in shaping mortality risk (*19-22*), we provide a more complete picture of the myriad biological mechanisms and disease pathways associated with each of these exposures in the UK Biobank.

A number of findings ask for further research. For example, we observed that while frequency of alcohol intake increased the risk of mortality, chronic liver disease, COPD/emphysema, breast and lung cancer, and osteoporosis, the association was in the opposite direction as expected for diabetes, ischemic heart disease, and chronic kidney disease. Previous research shows that UK Biobank participants are likely to misreport alcohol consumption as a function of higher disease burden (*23*), and that Mendelian randomization (MR) may be needed to properly estimate its association with specific diseases (*24*). We observed intake of glucosamine supplements to be an influential exposure in explaining mortality variation (Fig. 3d), and to be associated with decreased risk of many cardiometabolic diseases. Glucosamine supplementation was also cross-sectionally associated with many cardiometabolic disease mechanisms in our study (lower glucose, hbA1c, and triglycerides; higher HDL and total cholesterol), however previous clinical trials have reported no effect of oral glucosamine on blood pressure, glucose, and lipids (*25*). Further mechanistic or MR studies will be needed to disentangle whether glucosamine intake has a causal effect on these outcomes or instead whether our results reflect the fact that those who take glucosamine are also likely to

have higher socioeconomic status and less likely to be sick or disabled at baseline, as shown in the glucosamine PheWAS (Online Materials; Supplementary File SF85).

There are several limitations to note for our analysis. First, despite our prospective study design, we are unable to make causal conclusions based on our findings. Although consistent association patterns observed across different layers of outcomes (mortality, biomarkers, cardiometabolic risk factors, incident diseases) suggest that many exposures may play early roles in shaping age-related disease and mortality risk, causality will need to be verified with further methods such as MR. Second, exposome influences are dynamic over time and our study design cannot capture this dynamic aspect of the exposome since all exposures were only measured at one time point in the full cohort. We also have not captured all exposome influences, as we were limited to the exposures available in the UK Biobank. Conspicuously absent from our analysis are specific chemical and toxicological exposures beyond air pollution. Furthermore, a great deal of the exposure variables come from self-reported questionnaire data, which introduces potential recall bias, mismeasurement, and uncertainty into the reliability and accuracy of the responses. Finally, the use of existing PRS as proxies for the inherited genetic component of each disease is somewhat preliminary, as these are still being updated and improved and do not include the component of rare variation in single genes such as BRCA1/2 or the genes for familial hypercholesterolemia.

Despite these limitations, we believe that our approach offers many advantages over traditional single exposure approaches in environmental epidemiology. Through the use of: (i) independent discovery, replication, and validation stages; (ii) exposome-wide significance thresholds; (iii) sensitivity analyses to test for reverse causation bias; (iv) a comprehensively phenotyped cohort in which we could systematically conduct PheWAS of replicated exposures to test for mismeasurement in self-report variables; and (v) within-cluster and exposome-wide multivariable models to test for correlation bias, our approach greatly increases the reproducibility and positive predictive value of findings.

Overall, our results indicate that environment-focused interventions are likely to have the highest impact on ameliorating premature mortality and most age-related morbidity. The group of diseases with a small observed impact of the exposome relative to the genome are neurodegenerative disorders (AD, Parkinson's, macular degeneration) and some cancers (breast, prostate and colorectal). We argue that greater use of exposome study designs will significantly accelerate identification of high-priority population health targets for age-related morbidity and premature mortality. Our study also opens the door for further targeted proteomic, metabolomic, or other 'omics studies to explore the biological effects of the exposures that we identified in our study.

Figures

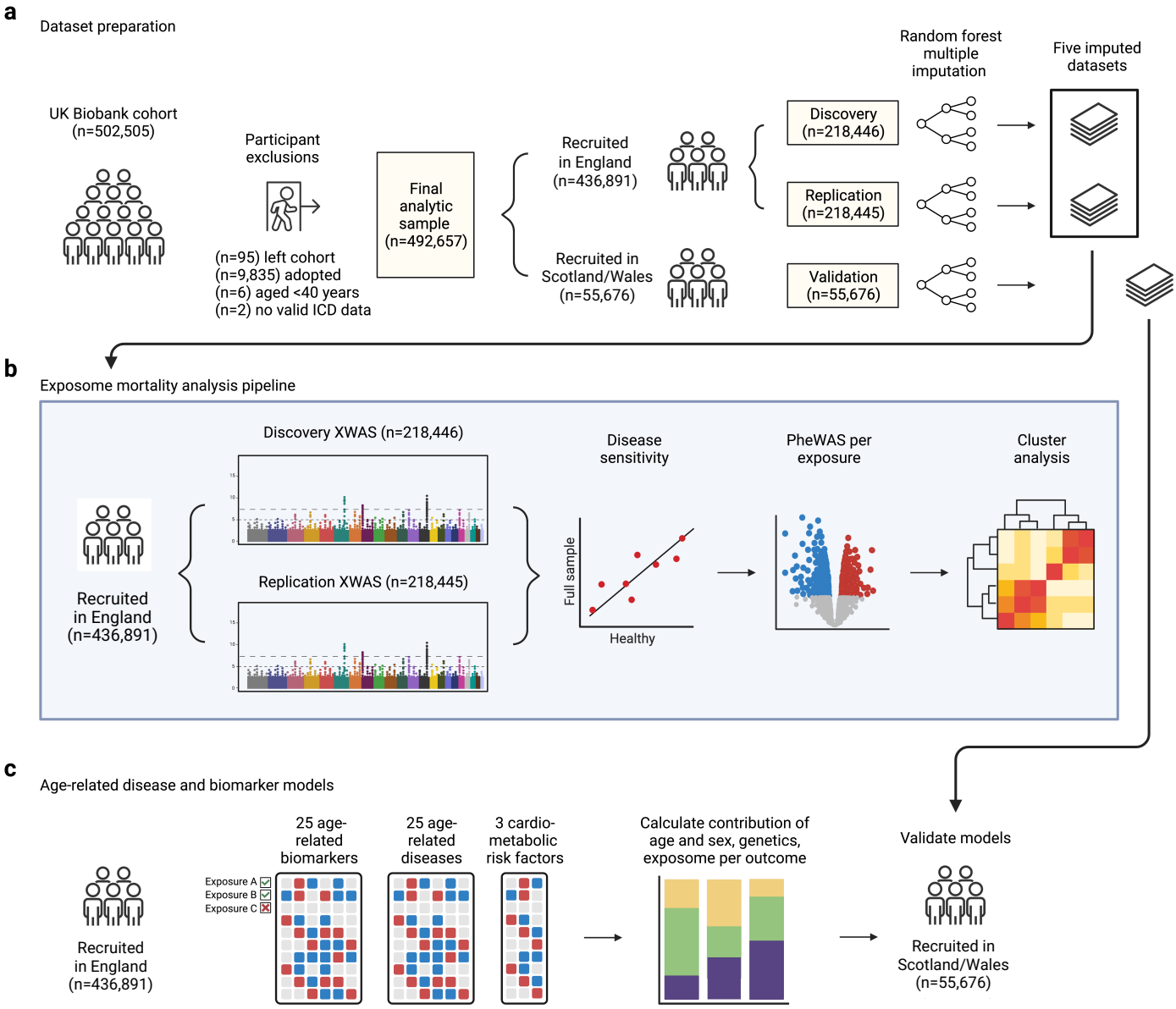

Fig. 1. Study overview. **(a)** After participant exclusions, UK Biobank participants were split into independent discovery, replication, and validation sets. Missing values were imputed separately within each group using random forest multiple imputation, resulting in 5 imputed datasets for each dataset. **(b)** Among UK Biobank participants recruited in England (n=436,891), an exposome-wide association study (XWAS) for all-cause mortality was conducted using the discovery and replication sets. The discovery and replication sets were then pooled, and further analyses were conducted in the full English sample to identify and remove replicated exposures that are sensitive to reverse causation (disease sensitivity), mismeasurement (PheWAS per exposure), and correlation bias (cluster analysis). **(c)** Exposures surviving all analyses in **(b)** were then tested in relation to 25 age-related biomarkers, 25 age-related diseases, and 3 cardiometabolic risk factors (hypertension, obesity, dyslipidemia). For mortality and each age-related disease, the relative contributions of age and sex, polygenic risk, and exposome were calculated via multivariable Cox proportional hazards models. Multivariable models were validated in participants recruited in Scotland/Wales (n=55,676), who were held out from all other analyses.

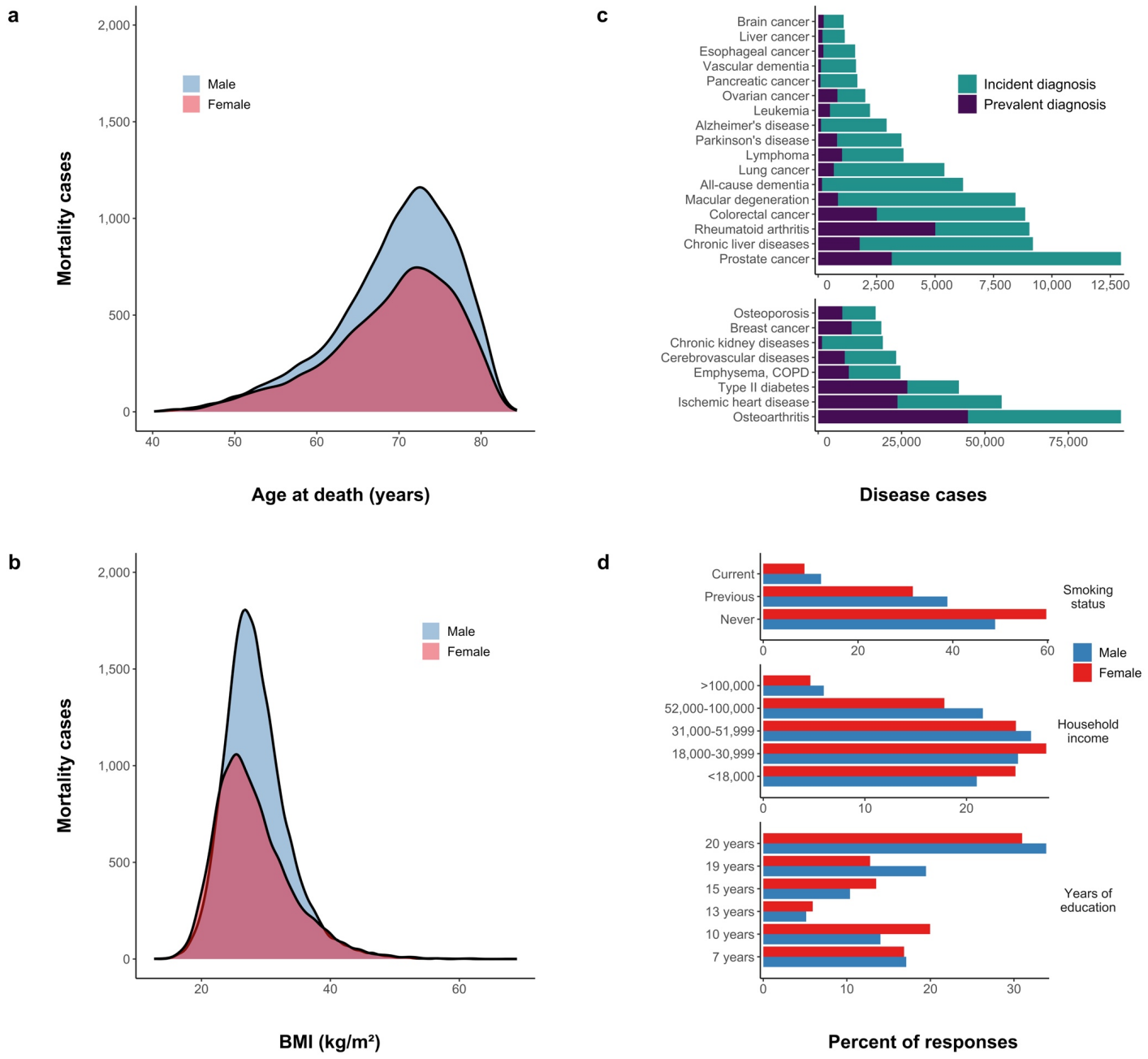

Fig. 2. Mortality and key exposure response rates among UK Biobank participants. **(a)** The number of deaths in women and men according to age at death (in years). **(b)** The number of deaths in women and men according to body mass index (BMI) at baseline. **(c)** Numbers of prevalent and incident cases for all age-related diseases studied. Note that diseases are put into two groups with different x-axis scales, since some diseases had far more cases than others. **(d)** Response rates for key exposures and covariates in both women and men. Percentages are for each sex separately and not across both sexes. All descriptive statistics are for UK Biobank participants recruited in England (n=436,891).

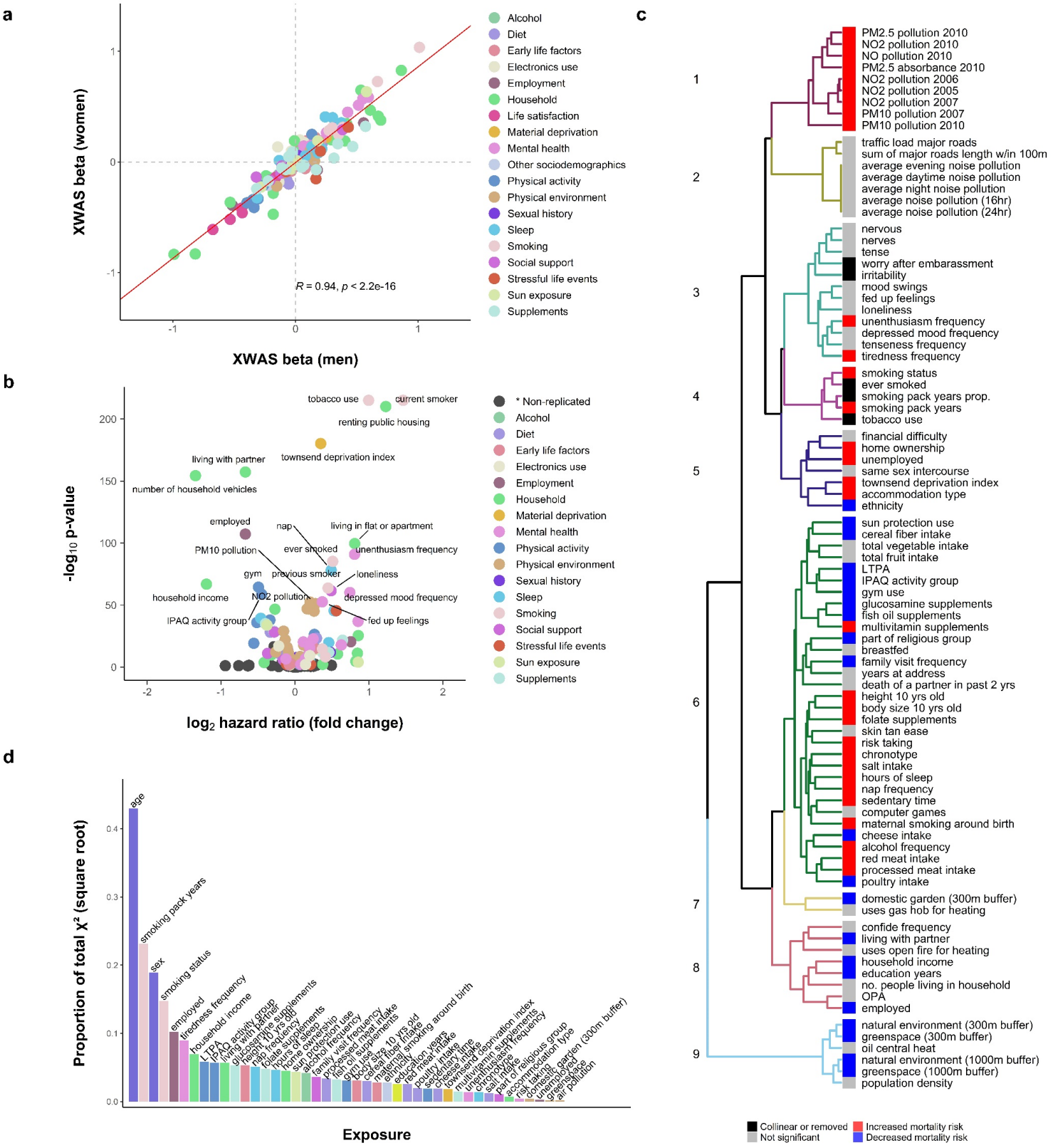
a
XWAS beta (women)
XWAS beta (men)
R = 0.94, p < 2.2e-16
Alcohol
Diet
Early life factors
Electronics use
Employment
Household
Life satisfaction
Material deprivation
Mental health
Other sociodemographics
Physical activity
Physical environment
Sexual history
Sleep
Smoking
Social support
Stressful life events
Sun exposure
Supplements
b
-log10 p-value
log2 hazard ratio (fold change)
tobacco use
current smoker
renting public housing
townsend deprivation index
living with partner
number of household vehicles
employed
nap
living in flat or apartment
PM10 pollution
ever smoked
unenthusiasm frequency
gym
previous smoker
loneliness
household income
NO2 pollution
depressed mood frequency
IPAQ activity group
fed up feelings
* Non-replicated
c
PM2.5 pollution 2010
NO2 pollution 2010
NO pollution 2010
PM2.5 absorbance 2010
NO2 pollution 2006
NO2 pollution 2005
NO2 pollution 2007
PM10 pollution 2007
PM10 pollution 2010
traffic load major roads
sum of major roads length w/in 100m
average evening noise pollution
average daytime noise pollution
average night noise pollution
average noise pollution (16hr)
average noise pollution (24hr)
nervous
nerves
tense
worry after embarassment
irritability
mood swings
fed up feelings
loneliness
unenthusiasm frequency
depressed mood frequency
tenseness frequency
tiredness frequency
smoking status
ever smoked
smoking pack years prop.
smoking pack years
tobacco use
financial difficulty
home ownership
unemployed
same sex intercourse
townsend deprivation index
accommodation type
ethnicity
sun protection use
cereal fiber intake
total vegetable intake
total fruit intake
LTPA
IPAQ activity group
gym use
glucosamine supplements
fish oil supplements
multivitamin supplements
part of religious group
breastfed
family visit frequency
years at address
death of a partner in past 2 yrs
height 10 yrs old
body size 10 yrs old
folate supplements
skin tan ease
risk taking
chronotype
salt intake
hours of sleep
nap frequency
sedentary time
computer games
maternal smoking around birth
cheese intake
alcohol frequency
red meat intake
processed meat intake
poultry intake
domestic garden (300m buffer)
uses gas hob for heating
confide frequency
living with partner
uses open fire for heating
household income
education years
no. people living in household
OPA
employed
natural environment (300m buffer)
greenspace (300m buffer)
oil central heat
natural environment (1000m buffer)
greenspace (1000m buffer)
population density
Collinear or removed
Not significant
Increased mortality risk
Decreased mortality risk
d
Proportion of total χ² (square root)
Exposure
age
smoking pack years
sex
smoking status
employed
tiredness frequency
household income
LTPA
IPAQ activity group
living with partner
glucosamine supplements
height 10 yrs old
nap frequency
folate supplements
hours of sleep
home ownership
sun protection use
alcohol frequency
family visit frequency
processed meat intake
fish oil supplements
gym use
body size 10 yrs old
cereal fiber intake
maternal smoking around birth
ethnicity
education years
red meat intake
poultry intake
sedentary time
cheese intake
townsend deprivation index
multivitamin supplements
unenthusiasm frequency
chronotype
salt intake
part of religious group
accommodation type
risk taking
domestic garden (300m buffer)
unemployed
greenspace
air pollution

Fig. 3. Environmental architecture of mortality in the UK Biobank. **(a)** Correlation between betas for the association between each exposure and mortality calculated separately in women (n=237,637) and men (n=199,257). **(b)** Volcano plot of log-transformed p-values and fold change (calculated as $\log_2$ of the hazard ratio) for all mortality XWAS associations in the final pooled analysis. Each point represents the effect and p-value for the association between a single exposure and all-cause mortality from the XWAS discovery analysis (n=218,483). Exposures that were FDR significant in both the discovery and replication stages are colored, whereas associations that were not replicated are colored dark grey and grouped in the category "* Non-replicated." The top 20 points according to strongest p-value are labelled. **(c)** Cluster structure of exposures that were replicated in the mortality XWAS and not discarded during reverse causation and PheWAS sensitivity analyses. Heatmap along the cluster dendrogram shows the effect direction and significance of each exposure in the final cluster multivariable model, with exposures that were not significant at the level of $p<0.05$ in the cluster multivariate models colored grey. Note that while individual exposures are shown, exposures in clusters 1, 2, and the greenspace/natural environment exposures in cluster 9 were reduced to principal components for multivariable modeling due to extremely high correlation (> 0.90) among similar exposures. **(d)** Individual exposure importance from a multivariable model including age, sex, and all 41 exposures identified in cluster modeling (n=436,891). Variable importance was calculated using a Wald test from ANOVA, where the importance of each variable is the proportion of that variable's Wald chi-squared (X^2) relative to the total model X^2. Note: the y-axis values were transformed by taking the square root to improve visualization. IPAQ: International Physical Activity Questionnaires; LTPA: leisure time physical activity; MH: mental health; OPA: occupational physical activity.

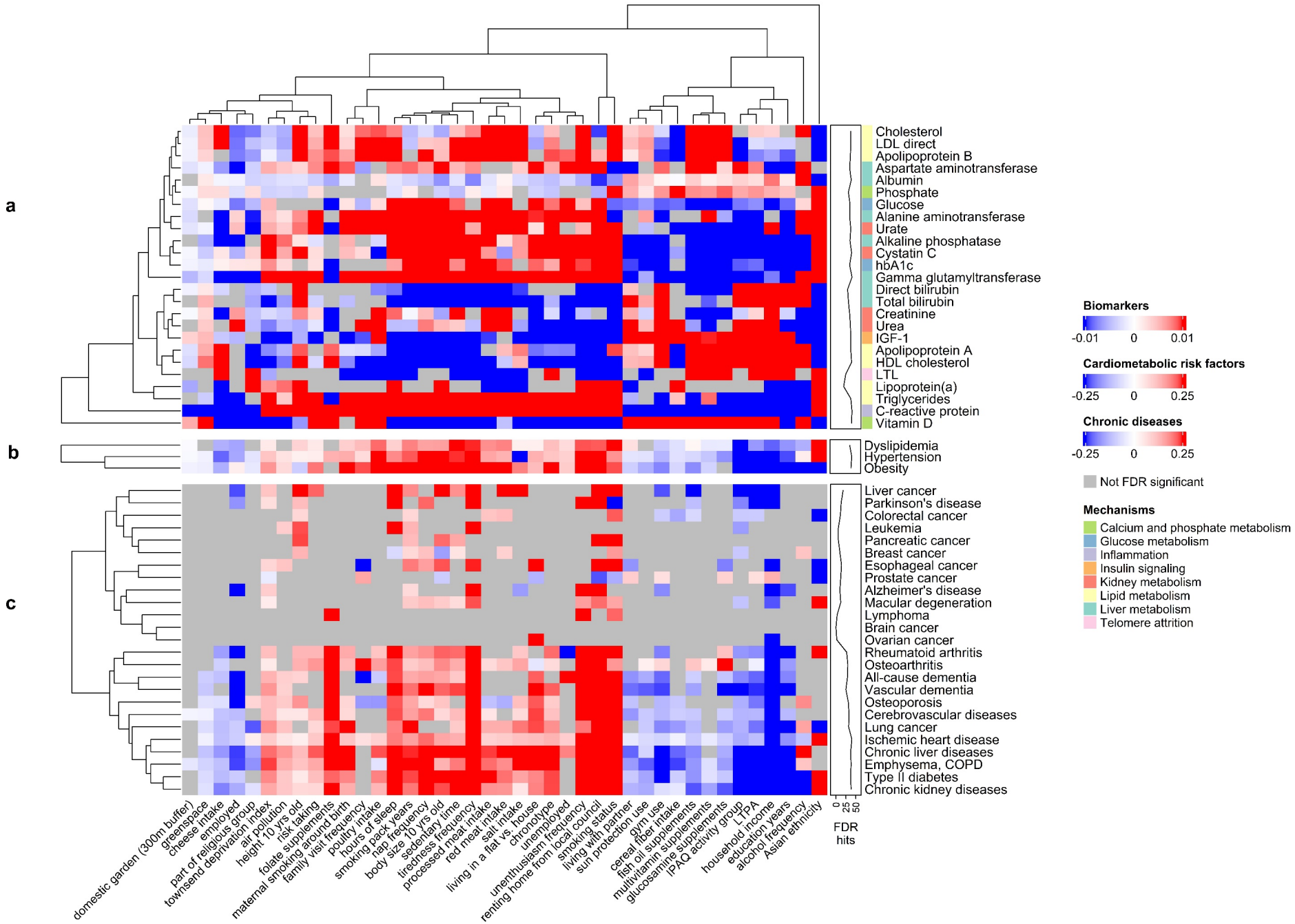
a
b
c
Cholesterol
LDL direct
Apolipoprotein B
Aspartate aminotransferase
Albumin
Phosphate
Glucose
Alanine aminotransferase
Urate
Alkaline phosphatase
Cystatin C
hbA1c
Gamma glutamyltransferase
Direct bilirubin
Total bilirubin
Creatinine
Urea
IGF-1
Apolipoprotein A
HDL cholesterol
LTL
Lipoprotein(a)
Triglycerides
C-reactive protein
Vitamin D
Dyslipidemia
Hypertension
Obesity
Liver cancer
Parkinson's disease
Colorectal cancer
Leukemia
Pancreatic cancer
Breast cancer
Esophageal cancer
Prostate cancer
Alzheimer's disease
Macular degeneration
Lymphoma
Brain cancer
Ovarian cancer
Rheumatoid arthritis
Osteoarthritis
All-cause dementia
Vascular dementia
Osteoporosis
Cerebrovascular diseases
Lung cancer
Ischemic heart disease
Chronic liver diseases
Emphysema, COPD
Type II diabetes
Chronic kidney diseases
domestic garden (300m buffer)
greenspace
cheese intake
employed
part of religious group
townsend deprivation index
air pollution
height 10 yrs old
risk taking
folate supplements
maternal smoking around birth
family visit frequency
poultry intake
hours of sleep
smoking pack years
nap frequency
body size 10 yrs old
sedentary time
tiredness frequency
processed meat intake
red meat intake
salt intake
living in a flat vs. house
chronotype
unemployed
unenthusiasm frequency
renting home from local council
smoking status
living with partner
sun protection use
gym use
cereal fiber intake
fish oil supplements
multivitamin supplements
glucosamine supplements
IPAQ activity group
LTPA
household income
education years
alcohol frequency
Asian ethnicity
0
25
50
FDR hits
Biomarkers
-0.01
0
0.01
Cardiometabolic risk factors
-0.25
0
0.25
Chronic diseases
-0.25
0
0.25
Not FDR significant
Mechanisms
Calcium and phosphate metabolism
Glucose metabolism
Inflammation
Insulin signaling
Kidney metabolism
Lipid metabolism
Liver metabolism
Telomere attrition

Fig. 4. Environmental architectures of age-related biological mechanisms and diseases in the UK Biobank. **(a)** Associations between each mortality-associated exposure and aging biomarkers. **(b)** Associations between each mortality-associated exposure and cardiometabolic risk factors. **(c)** Associations between each mortality-associated exposure and age-related chronic diseases. Colors in the heatmaps represent betas for associations between exposures and biomarkers/diseases. A line annotation track is shown that counts the total number of FDR significant associations for each outcome. For heatmap **(a)**, an addition annotation track shows the primary biological mechanism associated with each aging biomarker. Air pollution and greenspace are the first principal components combining all air pollution and greenspace variables, respectively. For nominal categorical variables with more than one response level, the association for the level with the strongest p-value is reported in this figure and the exposure's label reflects the response category shown. COPD: chronic obstructive pulmonary diseases; IGF-1: insulin-like growth factor 1; LTL: leucocyte telomere length; LTPA: leisure time physical activity.

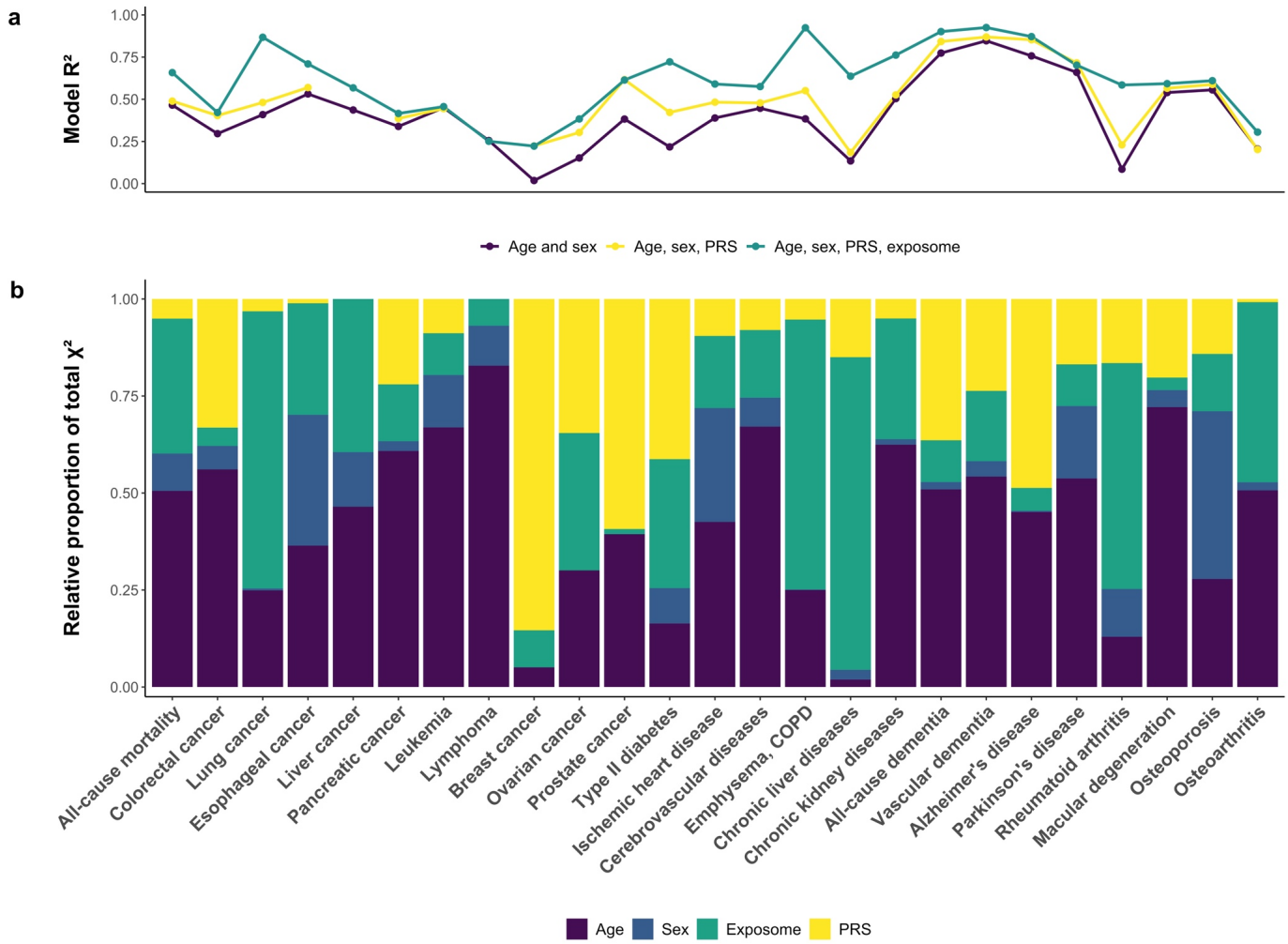

Fig. 5. Combined environmental and genetic architectures of mortality and age-related diseases. **(a)** R^2 calculated across studied outcomes for several sequential multivariable models: model 1 containing age and sex (purple); model 2 containing age, sex, and polygenic risk scores (PRS; yellow); and model 4 containing age, sex, PRS, and exposome (green). If a PRS was not available for a particular outcome, then the green R^2 shows the results from model 3 (age, sex, exposome). R^2 values are shown from the validation analyses in Scottish/Welsh validation set (n=55,676). **(b)** Variable importance for age, sex, polygenic risk, and exposome for all outcomes studied in model 4 conducted among English UK Biobank participants (n=436,891). Variable importance is calculated as the proportional of the total model chi-squared (X^2) that each variable category explains and is plotted as the relative contribution to the total model X^2 for each category so that they sum to 1. In all analyses, PRS includes polygenic risk scores, as well as genetic principal components and genotyping batch. PRS used for mortality models includes PRS for all other diseases and phenotypes shown in this figure (22 total). Note: PRS information was not available for liver cancer or lymphoma and is not included in models. Ovarian, breast, and prostate cancer models were sex-specific and so sex was not included in model 4 for these outcomes.

Materials and methods summary

Study design and participants

The UK Biobank is a prospective cohort study with extensive genetic and phenotype data available for 502,505 individuals resident in the United Kingdom (*26*). The full UK Biobank protocol is available online.

Exposures

We considered as potential exposome-wide association study (XWAS) exposures all non-genetic variables available as of July 24, 2020 that were collected at baseline and were available for participants recruited across all assessment centers. After all exclusions, recoding, and quality control (Supplementary Information), 176 unique exposures remained that were available in the full cohort that were common to both women and men. All continuous exposure variables were centered and standardized before analysis, except for age at recruitment. All ordinal categorical variables were recoded to only test linear associations and other polynomial contrasts (e.g., quadratic or cubic associations) were not assessed. All nominal categorical exposures were analyzed with the most common category set as the reference. Detailed data dictionaries including all exposures used in imputation and XWAS steps are included in Supplementary Files SF1-SF2.

Outcomes

Detailed information about the linkage procedure with national registries for mortality and cause of death information is available online. Mortality data were accessed from the UK Biobank data portal on May 4, 2022, with a censoring date of September 30, 2021 or October 31, 2021 for participants recruited in England/Scotland or Wales, respectively (11-15 years of follow-up).

Aging biomarkers (Table S8) were measured using baseline non-fasting blood serum samples as previously described (*27*). Data on leukocyte telomere length (LTL) was only available in a slightly smaller sample (n=472,506) than other biomarkers and was not imputed. Biomarkers were previously adjusted for technical variation by the UK Biobank, with sample processing and quality control procedures described on the UK Biobank website.

Data used to define chronic diseases and cardiometabolic risk factors are outlined in Table S9. Incident chronic disease diagnoses were ascertained using ICD diagnosis codes and corresponding dates of diagnosis taken from linked hospital inpatient records and death register data. ICD data were accessed from the UK Biobank data portal on May 30, 2022, with a censoring date of September 30, 2021; July 31, 2021; or February 28, 2018 for participants recruited in England, Scotland, or Wales, respectively (8-15 years of follow-up).

Missing data imputation

UK Biobank participants recruited from England were randomly assigned to a discovery (n=218,446) or replication set (n=218,445) while maintaining the same proportion of mortality cases in each. We performed missing data imputation separately in the discovery, replication, and Scottish/Welsh validation (n=55,676) datasets using the R package missRanger (*28*), which combines random forest imputation with predictive mean matching (Supplementary

Information). We imputed 5 datasets, with a maximum of 10 iterations for each imputation. All subsequent study analyses were run independently in each of the five imputed datasets, and results were pooled using Rubin's rule (*29*).

Exposome-wide association study

XWAS of all-cause mortality were initially carried out separately in women and men, and then a final XWAS was calculated in the pooled dataset with both women and men to increase power. Exposures in the final pooled XWAS were limited to those asked to both women and men, omitting sex-specific reproductive factors. In each XWAS, we serially assessed associations of each individual exposure with all-cause mortality using Cox proportional hazards models with age as the time scale; stratified by 5-year birth cohorts and sex (in the pooled analysis only); and adjusted for assessment center, years of education (7 years, 10 years, 13 years, 15 years, 19 years, 20 years), and ethnicity (White, Asian, Black, Mixed, Other). Since it has been shown that UK Biobank participants are likely to misreport alcohol consumption as a function of higher disease burden (*23*), self-reported overall health status was added as an additional XWAS covariate for the self-reported alcohol intake exposure only. P-values in the discovery and replication analyses were corrected using the false discovery rate (FDR; Benjamini-Hochberg method(*30*)) with a significance threshold of FDR $p < 0.05$. After completing the mortality XWAS, discovery and replication sets were recombined into the full English sample (n= 436,891) to complete further sensitivity analyses (Supplementary Information).

Correlation and cluster analyses

Correlation between all variables was calculated in the full English sample using the R package polycor (*31*) to create a heterogenous correlation matrix for each imputed dataset (Supplementary Information). We used hierarchical clustering via Euclidean distance to identify the cluster structure of exposures replicated in the pooled XWAS and not susceptible to reverse causation bias (plus education and ethnicity). Within-cluster Cox multivariable mortality models included all remaining variables in the cluster after removing collinear variables (Supplementary Information), with additional adjustment for assessment center, household income (Less than 18,000, 18,000 to 30,999, 31,000 to 51,999, 52,000 to 100,000, Greater than 100,000), education, and ethnicity (if those variables were not already in the cluster). As in the XWAS, additional adjustment for self-reported overall health status was made in the cluster with self-reported alcohol intake. The significance threshold used in cluster multivariable analyses was a nominal $p < 0.05$.

Aging mechanisms and incident chronic disease analyses

Aging biomarker variables were log transformed and then were age-adjusted by regressing each onto age at recruitment separately in women and men. Across exposures replicated in the XWAS and passing all sensitivity tests, we serially assessed associations between each exposure and age-adjusted biomarker using cross-sectional linear regression models with covariates for sex, 5-year birth cohort, assessment center, years of education, ethnicity, number of medications, smoking status (current, previous, never), and IPAQ physical activity level (low, moderate, high). IGF-1, LTL, and vitamin D models included additional covariates for standing height (in cm), leukocyte count (10^9 cells/Liter), and month of biomarker assessment (to control for seasonality of sun exposure), respectively.

For chronic disease analyses, we serially assessed associations between each exposure and incident disease using the Cox proportional hazards model, with all XWAS covariates plus household income, smoking status, and IPAQ physical activity group. Sex-specific reproductive exposures (e.g., menopause) replicated in the female- and male-only XWAS analyses were also tested as exposures in analyses of sex-specific chronic disease outcomes (breast, ovarian, and prostate cancer).

For cardiometabolic risk factors (obesity, hypertension, dyslipidemia), we serially assessed each exposure and risk factor pair using cross-sectional logistic regression models adjusted for age, sex, assessment center, household income, years of education, ethnicity, smoking status, and IPAQ physical activity level.

Across all biomarker, chronic disease, and cardiometabolic risk factor analyses, p-values were corrected separately for each outcome using FDR.

Exposome and polygenic risk multivariable models

For each outcome, five multivariable models were calculated. The first only includes age (scaled) and sex in the model (model 1). Model 2 includes age, sex, and the polygenic risk score (PRS) for the outcome, if available (see below for more detail). Model 3 includes age, sex, and all exposures associated with the outcome (exposome). Model 4 includes age, sex, exposome, and PRS. If a PRS was not available for a particular outcome, then models 2 and 4 were not calculated for that outcome. Each model was validated in the independent Scottish/Welsh dataset (n=55,676) by obtaining the linear predicted values from the models in the English dataset and measuring the C-index and R^2 for these values in relation to the outcome rates in the Scottish/Welsh population.

In all multivariable Cox models, the proportional hazards assumption was tested by examining the Schoenfeld residuals, and an interaction with time was added to any variable with

non-proportional hazards (Supplementary Information). Relative importance for each variable and category of variables within the multivariable models was calculated using Wald chi-squared (X^2) statistics via ANOVA, where the relative importance of each is the proportion of the variable/group X^2 relative to the total model X^2.

References and Notes

1. D. W. Belsky *et al.*, Quantification of biological aging in young adults. *Proc Natl Acad Sci U S A* **112**, E4104-4110 (2015).
2. C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, G. Kroemer, The Hallmarks of Aging. *Cell* **153**, 1194-1217 (2013).
3. H. S. Chahal, W. M. Drake, The endocrine system and ageing. *J Pathol* **211**, 173-180 (2007).
4. C. Franceschi *et al.*, Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann N Y Acad Sci* **908**, 244-254 (2000).
5. C. Franceschi, P. Garagnani, P. Parini, C. Giuliani, A. Santoro, Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol* **14**, 576-590 (2018).
6. L. Partridge, J. Deelen, P. E. Slagboom, Facing up to the global challenges of ageing. *Nature* **561**, 45-56 (2018).
7. E. Fabbri *et al.*, Aging and Multimorbidity: New Tasks, Priorities, and Frontiers for Integrated Gerontological and Clinical Research. *J Am Med Dir Assoc* **16**, 640-647 (2015).
8. A. Kingston, L. Robinson, H. Booth, M. Knapp, C. Jagger, Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. *Age and Ageing* **47**, 374-380 (2021).
9. G. A. Collaborators, Global, regional, and national burden of diseases and injuries for adults 70 years and older: systematic analysis for the Global Burden of Disease 2019 Study. *BMJ* **376**, e068208 (2022).
10. J. G. Ruby *et al.*, Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating. *Genetics* **210**, 1109-1124 (2018).
11. J. Kaplanis *et al.*, Quantitative analysis of population-scale family trees with millions of relatives. *Science* **360**, 171-175 (2018).
12. J. Oeppen, J. W. Vaupel, Demography. Broken limits to life expectancy. *Science* **296**, 1029-1031 (2002).
13. J. C. Riley, *Rising life expectancy: a global history*. (Cambridge University Press, Cambridge, 2001), pp. xii, 243 p.
14. R. Vermeulen, E. L. Schymanski, A. L. Barabási, G. W. Miller, The exposome and health: where chemistry meets biology. *Science* **367**, 392-396 (2020).
15. V. Kalia, D. W. Belsky, A. A. Baccarelli, G. W. Miller, An exposomic framework to uncover environmental drivers of aging. *Exposome* **2**, osac002 (2022).
16. L. Broer *et al.*, Distinguishing true from false positives in genomic studies: p values. *Eur J Epidemiol* **28**, 131-138 (2013).
17. J. P. Ioannidis, R. Tarone, J. K. McLaughlin, The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* **22**, 450-456 (2011).
18. C. J. Patel *et al.*, Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States national health and nutrition examination survey. *Int J Epidemiol* **42**, 1795-1810 (2013).
19. S. Stringhini *et al.*, Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet* **389**, 1229-1237 (2017).
20. K. Blond, C. F. Brinkløv, M. Ried-Larsen, A. Crippa, A. Grøntved, Association of high amounts of physical activity with mortality risk: a systematic review and meta-analysis. *Br J Sports Med* **54**, 1195-1201 (2020).

21. S. Yusuf *et al.*, Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* **395**, 795-808 (2020).
22. S. A. Lear *et al.*, The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the PURE study. *Lancet* **390**, 2643-2654 (2017).
23. A. Xue *et al.*, Genome-wide analyses of behavioural traits are subject to bias by misreports and longitudinal changes. *Nat Commun* **12**, 6450 (2021).
24. J. Lankester, D. Zanetti, E. Ingelsson, T. L. Assimes, Alcohol use and cardiometabolic risk in the UK Biobank: A Mendelian randomization study. *PLoS One* **16**, e0255801 (2021).
25. R. Palma Dos Reis *et al.*, Crystalline glucosamine sulfate in the treatment of osteoarthritis: evidence of long-term cardiovascular safety from clinical trials. *Open Rheumatol J* **5**, 69-77 (2011).
26. L. Palmer, UK Biobank: bank on it. *Lancet* **369**, 1980-1982 (2007).
27. P. Elliott, T. C. Peakman, The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology* **37**, 234-244 (2008).
28. M. Mayer. missRanger: Fast Imputation of Missing Values. R package version 2.1.0 (2019); https://CRAN.R-project.org/package=missRanger.
29. A. Marshall, D. Altman, R. Holder, P. Royston, Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* **9**, 57 (2009).
30. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57**, 289-300 (1995).
31. J. Fox. polycor: Polychoric and Polyserial Correlations. R package version 0.7-10 (2019); https://CRAN.R-project.org/package=polycor.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 61054. **Funding:** A.N-H. receives research funding from Novo Nordisk, GSK, and Ono Pharma. A.D. is supported by the Wellcome Trust [223100/Z/21/Z], Novo Nordisk, Swiss Re, the British Heart Foundation Centre of Research Excellence (grant number RE/18/3/34214), and Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. C.M.vD. is supported by the common mechanisms and pathways in Stroke and Alzheimer's disease (CoSTREAM) project (www.costream.eu, grant agreement No. 667375) and ZonMW Memorabel program (project number 733050814). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. **Author contributions:** M.A.A., C.M.vD., A.A., and N.A. designed the study. M.A.A. conducted all data preparation and analysis, with assistance from A.A. and N.A. Analyses were supervised by C.M.vD., A.A., and N.A. Analytical input was provided by W.S. and A.N-H. for clustering analyses. Analytic input for physical activity variables was provided by A.D. Guidance on calculating polygenic risk scores (PRS) was provided by J.C. for diseases not included in the UK Biobank PRS release (pancreatic, lung, esophageal, leukemia, COPD, chronic liver disease, chronic kidney disease, osteoarthritis) under the supervision of D.H. The systematic review was conducted by M.A.A. and S.M.K., including independent screening of abstracts/papers. M.A.A. prepared the manuscript, figures, tables, and supplementary files, with edits and revisions provided by all other co-authors. The GitHub code repository was created and is maintained by M.A.A. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** UK Biobank data are available through a procedure described at

https://www.ukbiobank.ac.uk/enable-your-research. R code needed to reproduce all analysis, figures, and tables is publicly available at https://github.com/miargentieri/exposome-aging-UK-Biobank. Summary statistics from all analysis stages are included in Supplementary Files SF3-SF108.

List of Supplementary Materials

Supplementary Information

Materials and Methods

Sensitivity Analyses

Supplementary discussion of findings in relation to existing research

Supplementary discussion on study limitations and strengths

Fig. S1 – S58

Tables S1 – S16

References (1 – 87)

Supplementary File titles and summaries

Supplementary File

This single Excel file contains Supplementary Files SF1-SF108; see main Supplementary Information PDF for table titles and summaries.

Online Materials

Interactive plots for XWAS and PheWAS analyses are shown on the paper's GitHub page: https://github.com/miargentieri/exposome-aging-UK-Biobank.